Poster Tag: WED-PM-200

RIFormer: Keep Your Vision Backbone Effective but Removing Token Mixer

Jiahao Wang[1,2], Songyang Zhang[1], Yong Liu[3], Taiqiang Wu[3], Yujiu, Yang[3], Xihui Liu[2], Kai Chen[1], Ping Luo[2], Dahua Lin[1]

[1]Shanghai AI Laboratory & [2]The Chinese University of HongKong & [3]Tsinghua Shenzhen International Graduate School, Tsinghua University

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory
香 港 大 學 THE UNIVERSITY OF HONG KONG
清華大学 Tsinghua University

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

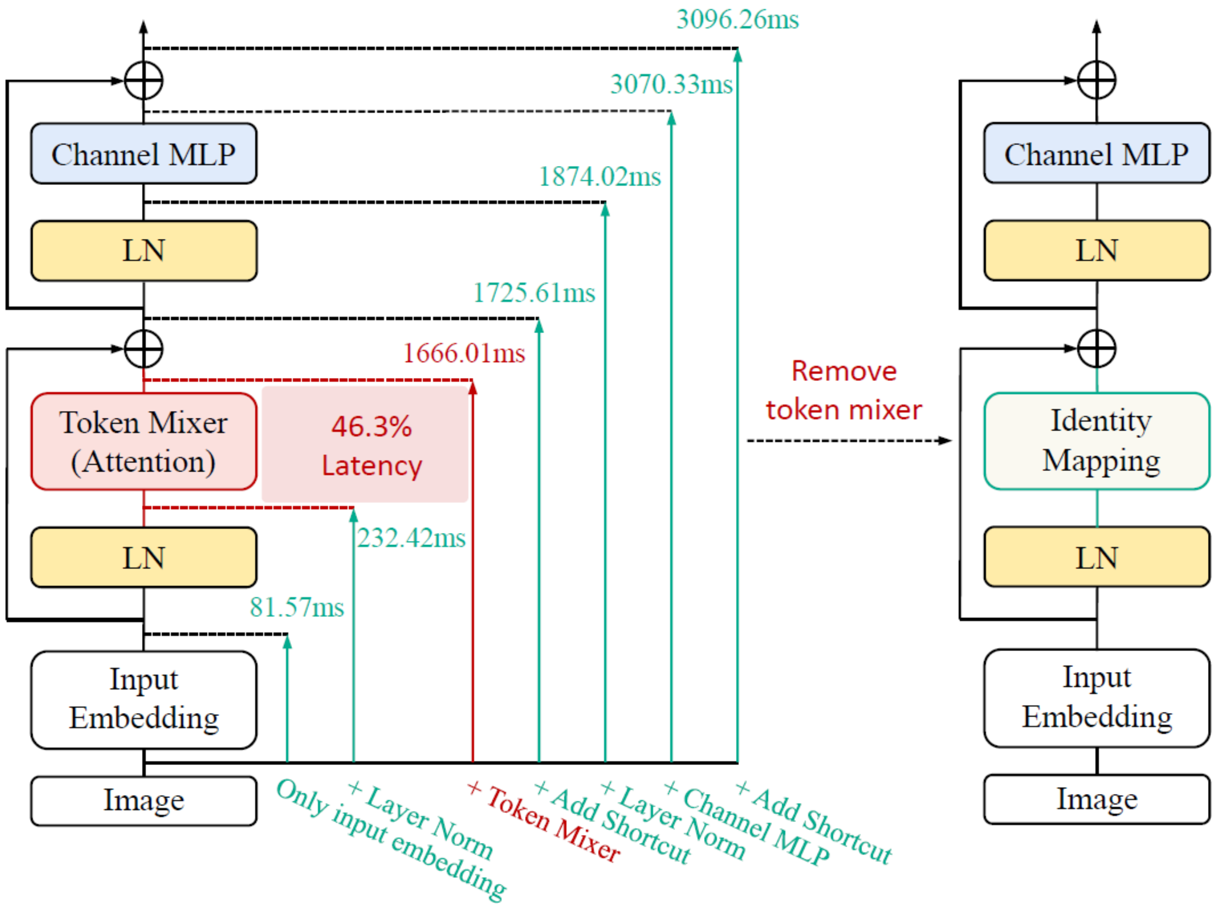
Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

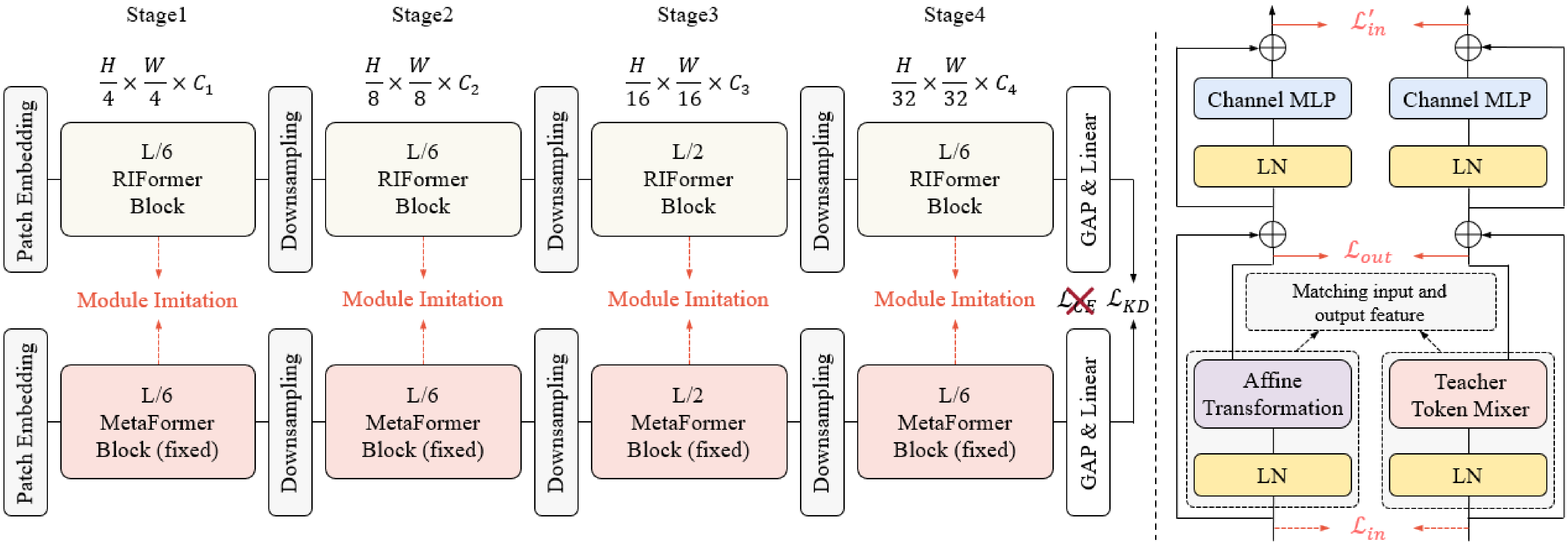
Overview

Summary

- **Token mixer in Vision Backbone:** Performing information communication between different spatial tokens but suffer from considerable computational cost and latency.
- **Efficient Foundation Model:** How to keep a vision backbone effective while removing token mixers.
- **RIFormer:** A token mixer-free model architecture.
- **Improved learning paradigm:** 5 practical guidelines.



Overview



Highlights

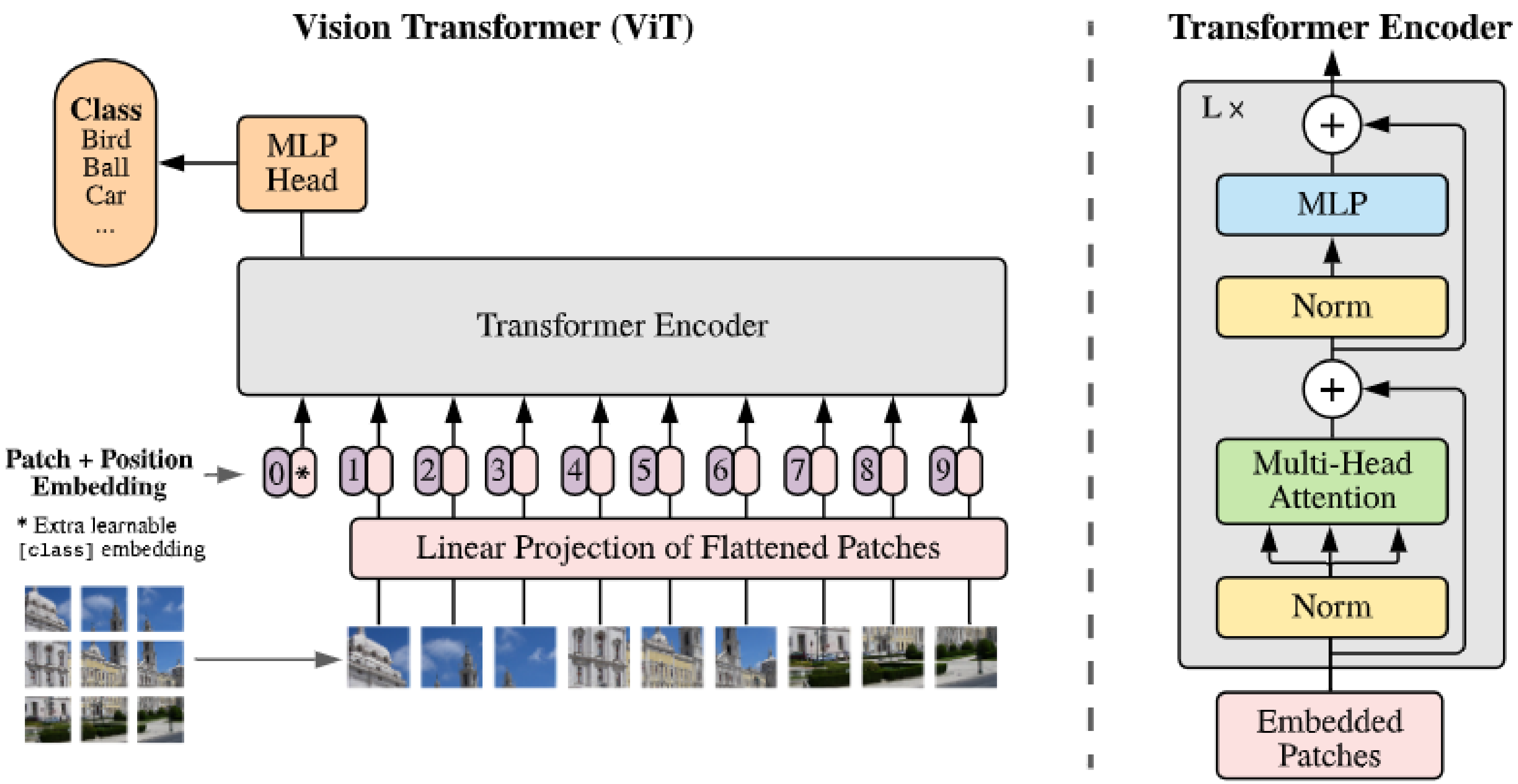
- The inductive bias of neural network, can be incorporated into simple network structure with appropriate optimization strategy.
- We hope this work can serve as a starting point for the exploration of optimization-driven efficient network design.

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

Background

ViT



Transformer Encoder as General Vision Backbone + Classification head for various downstream tasks

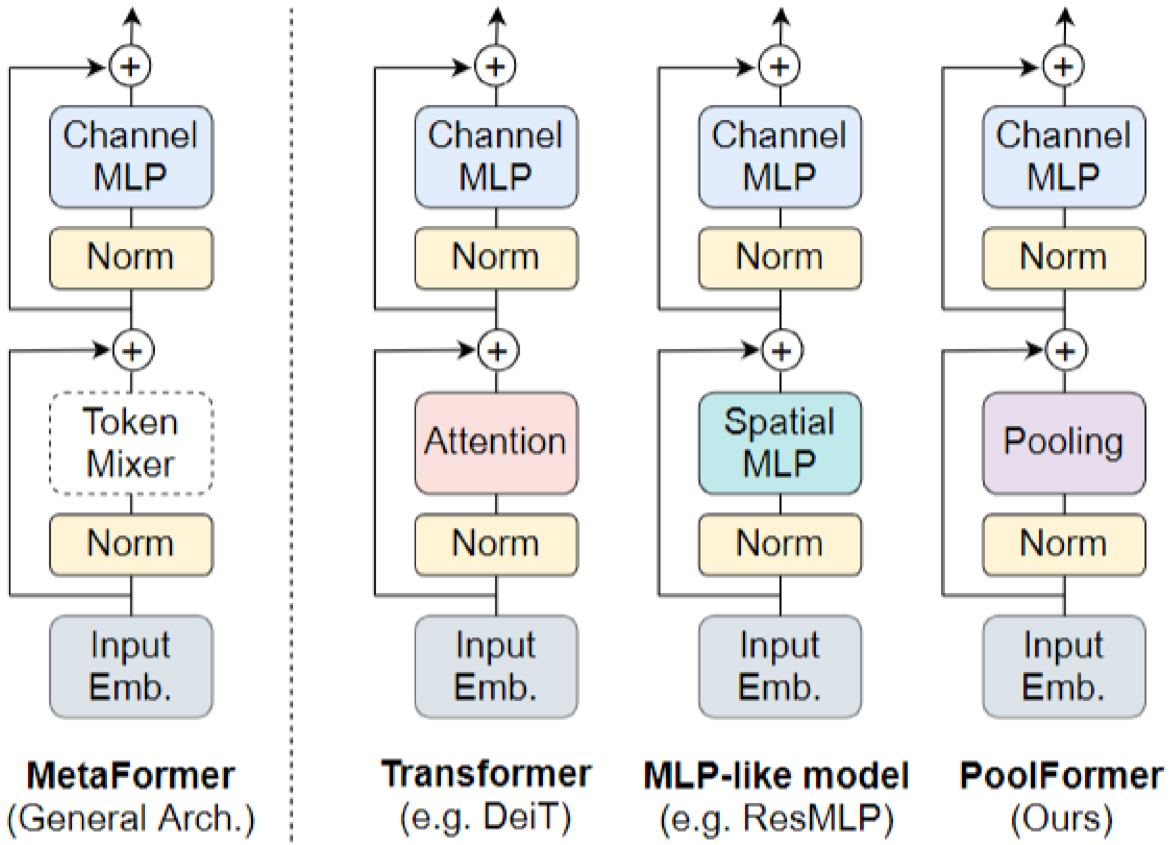
[1] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Background

MetaFormer: An Abstract Architecture of Transformer

- First sub-block: Token Mixer + LN
- Second sub-block: FFN + LN

- Token mixer is not specified to self-attention
- The other components are kept the same as Transformers



[1] MetaFormer Is Actually What You Need for Vision

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

Motivation

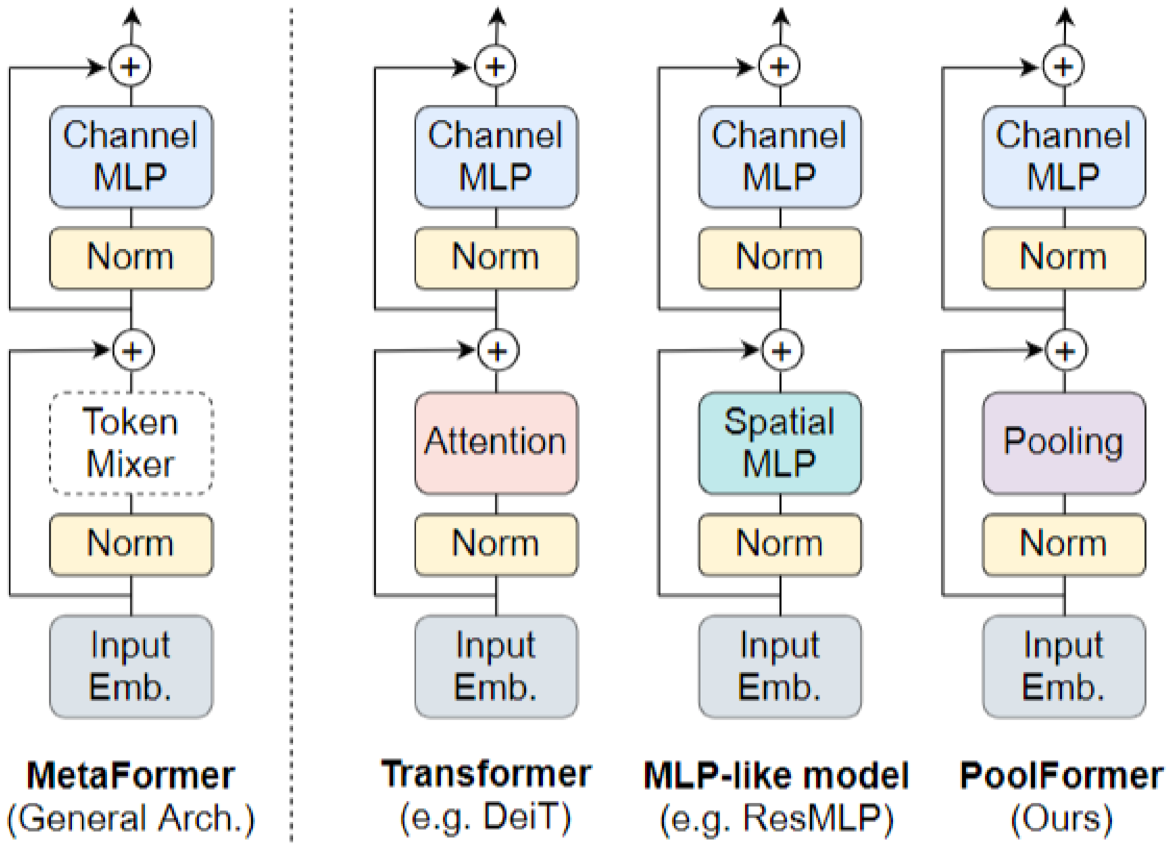
Token Mixer Can Be Simplified

Attention: Window-Based Attention[1], …

MLP: Spatial FC[2], CycleFC[3], AMixer[4], …

FFT: 2D FFT[5]

Others: Pooling[6]



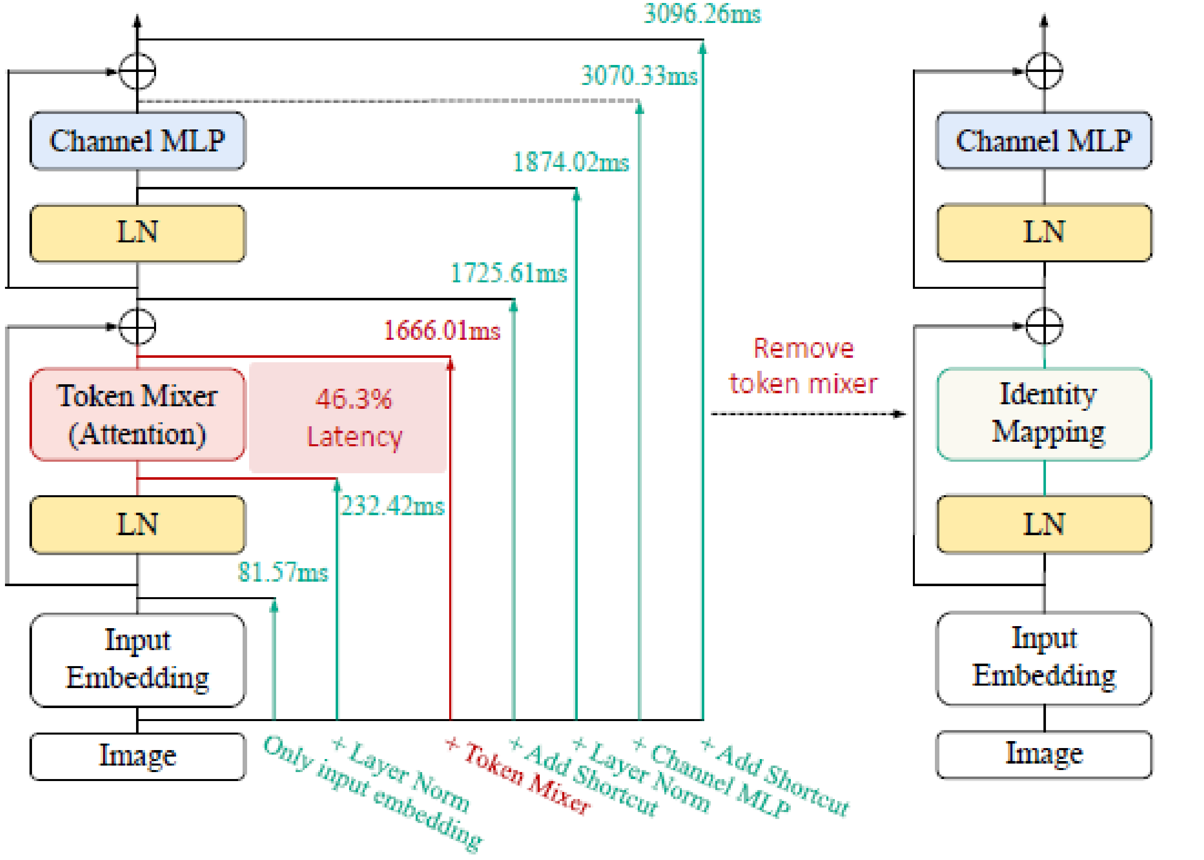
[1] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows
[2] ResMLP: Feedforward networks for image classification with data-efficient training
[3] CycleMLP: A MLP-Like Architecture for Dense Prediction
[4] AMixer: Adaptive Weight Mixing for Self-attention Free Vision Transformers
[5] Global Filter Networks for Image Classification
[6] MetaFormer Is Actually What You Need for Vision

Motivation

Can Token Mixer Be Completely Removed ?

- Token Mixer is heavy in latency
- For self-attention, the latency occupies about 46.3% of the backbone

Can we keep the vision backbone effective but removing the token mixer?



(a) Latency analysis of ViT-B (b) Remove token mixer with heavy latency

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

Advantage of Token Mixer-Free Architecture

- Embarrassingly Simple Architecture: 0 Params，0 FLOPs，0 Latency in token mixing, and reducing model latency, power consumption and memory usage

- Only one operation, Channel MLP (1×1 Conv): An inference chip specialized for RIFormer can have an enormous number of LN-1×1 units, facilitating hardware specialization to achieve even higher speed (the fewer types of operators we require, the more computing units we can integrate onto the chip[1])

- Decoupling the Complexity of the Model During Training and Inference:
- Training: Use affine transformation → Enhance representation capability
- Inference: Affine operator can be integrated into the previous LN → Completely cut off token mixer

- To the community: Focus more on the overall architecture and training strategy of ViTs-like models, rather than only on the design of the token mixer

[1] RepVGG: Making VGG-style ConvNets Great Again

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step1: Directly Train a Vision Backbone Without Token Mixer

- Baseline: PoolFormer[1]
- Dataset: ImageNet-1K
- Epoch: 120
- Optimizer: AdamW

Token Mixer	Training recipe	ImageNet top-1 acc (%)
Pooling	CE Loss	**75.01**
Identity	CE Loss	72.31

Trivial supervised training can lead to an unacceptable performance drop (2.7% top-1 accuracy)

We need more advanced training procedure

[1] MetaFormer Is Actually What You Need for Vision

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step2: Knowledge Distillation

Hard distillation or Soft distillation ?

- RIFormer shares the same macro structure as transformer
- Cannot be treated as a student transformer as no self-attention

- Do not prefer viewing it as a pure convnet
- Resemblance to transformer in terms of macro/micro-level architecture design

Use Cross Entropy Loss or not ?

- Label smoothing:
  hard label → $1-\varepsilon$ true label + ε uniform distribution

- 1×1 convolutions dominate basic building block in RIFormer, such a simplified design require richer information in the supervised labels

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step2: Knowledge Distillation

- Teacher: GFNet-H-B[1]
- Dataset: ImageNet-1K
- Epoch: 120
- Optimizer: AdamW

TM	Label	Teacher	ImageNet top-1 acc (%)
Identity	✓	✗	72.31
Identity	✓	hard	73.51
Identity	✗	hard	72.86
Identity	✓	soft	73.64
Identity	✗	soft	**74.05**

Supervised Training + Hard Distillation (DeiT[2]) does not seem to be the most suitable way for a crude model without token mixer

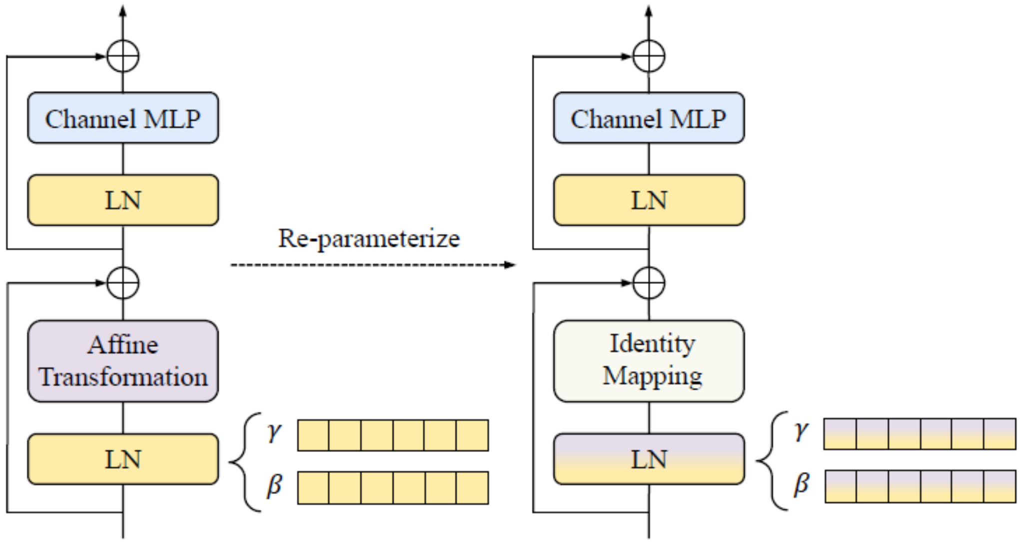
Pure Soft Distillation (Optimal) still fails to fully recover the performance gap

[1] Global Filter Networks for Image Classification
[2] Training data-efficient image transformers & distillation through attention

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step3: Structural Re-parameterization



(a) RepIdentityFormer Training (b) RepIdentityFormer Inference

Training-time module should satisfy:

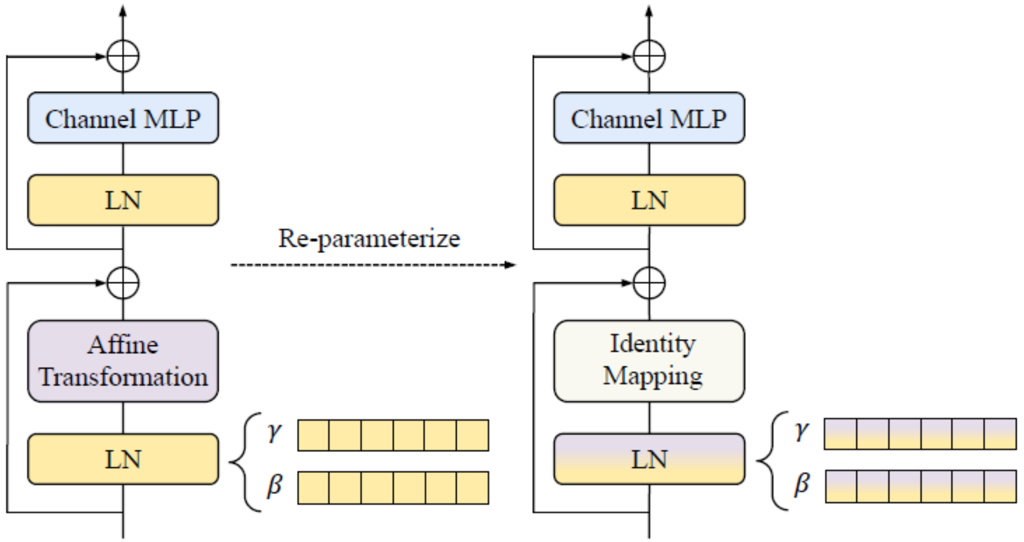
1. per-location operator for allowing equivalent transformation
2. parametric operator for allowing extra representation ability

We choose the affine transformation

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step3: Structural Re-parameterization



(a) RepIdentityFormer Training (b) RepIdentityFormer Inference

TM	Label	KD type	ImageNet top-1 acc (%)
Affine	✓	✗	72.25
Affine	✓	hard	73.44
Affine	✗	hard	72.77
Affine	✓	soft	72.10
Affine	✗	soft	**74.07**

Using affine transformation without tailored distillation, is hard to recover the performance degradation

The affine transformation in the LN is a linear transformation that can be directly merged with the extra affine operator we introduced

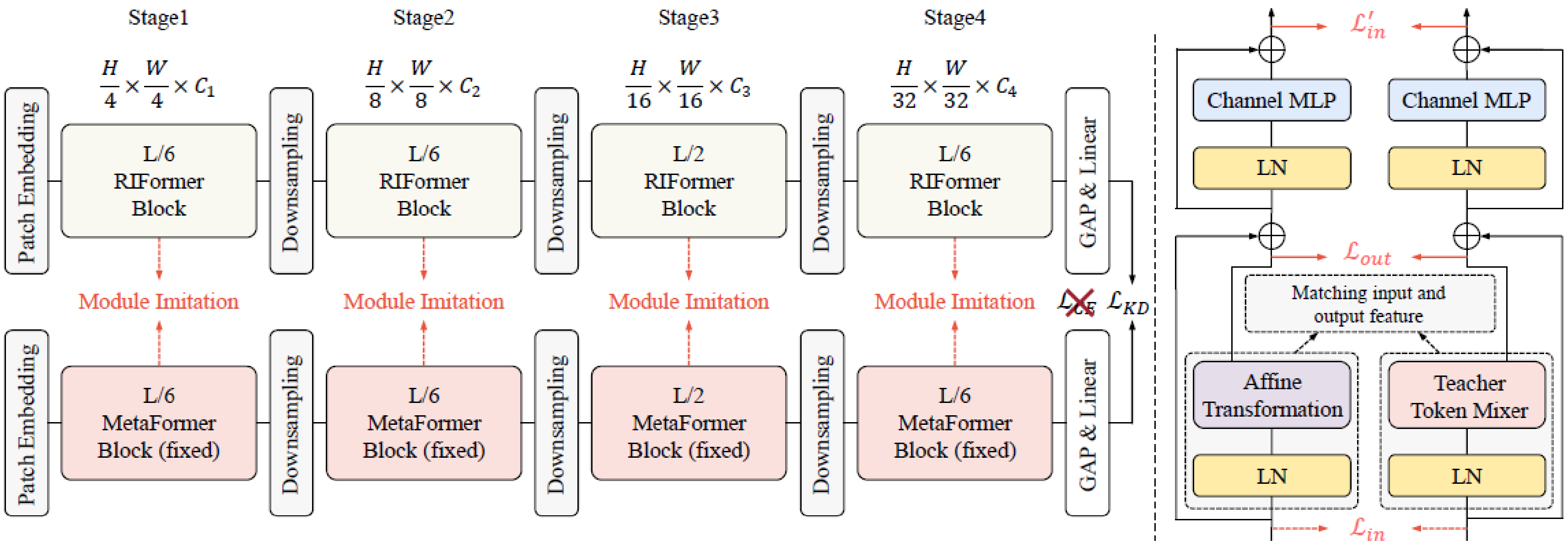
A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step4: Module Imitation: let the affine module to approximate the behavior of the token mixer module

$$\mathcal{L} = \mathcal{L}_{soft} + \lambda_1 \mathcal{L}'_{in} + \lambda_2 \mathcal{L}_{out} + \lambda_3 \mathcal{L}_{rel},$$

- Teacher (out): GFNet-H-B
- Teacher (feature): PoolFormer-S12
- Dataset: ImageNet-1K
- Epoch: 120



(a) Overall training procedure with L RIFormer blocks (b) Module imitation

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step4: Module Imitation: let the affine module to approximate the behavior of the token mixer module

- Module imitation, helps leveraging the modeling capacity of affine operator, by helping the affine operator implicitly benefit from the supervision of the teacher’ s token mixer

TM	Feat	Rel	Layer	ImageNet top-1 acc (%)
Affine	0	0	-	74.07
Affine	40	0	6	74.49
Affine	60	0	6	74.77
Affine	80	0	6	74.81
Affine	80	10	6	75.08
Affine	80	20	6	74.82
Affine	80	40	6	75.00
Affine	80	20	4	**75.13**

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step5: What Kind of Teacher is Better for the Token Mixer-Free Architecture ?

- Teacher with large receptive field is beneficial to improve student with limited receptive field
- Receptive field gap between teacher and student: inductive bias can be transferred from one model to another through distillation[1]

Teacher (T)	T.acc (%)	MI	ImageNet top-1 acc (%)
PoolFormer-M48 [52]	82.5	✗	73.63
Swin-B* [26]	85.2	✗	73.12
Pyramid ViG-B [15]	83.7	✗	73.25
GFNet-H-B [34]	82.9	✗	74.07
PoolFormer-M48 [52]	82.5	✓	74.83
Swin-B* [26]	85.2	✓	74.52
Pyramid ViG-B [15]	83.7	✓	74.25
GFNet-H-B [34]	82.9	✓	**75.13**

A Roadmap

From a Fully Supervised Approach to a More Advanced Paradigm

Step6: Load Partial Parameters From Teacher

- w/o loading the pre-trained weight of teacher model: 75.13%
- w/ loading the pre-trained weight of teacher model: 75.36%
- Dataset: ImageNet-1K
- Epoch: 120

- Load the pre-trained weight of teacher model (except the token mixer) into student improve the convergence and performance

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

Evaluation

ImageNet-1K Evaluation

- Favorable speed advantage is achieved
- RIFormer shows promising results
- Optimization strategy plays a key role

Token Mixer	Outcome Model	Image Size	Params (M)	MACs (G)	Throughput (images/s)	Top-1 (%)
Convolution	RSB-ResNet-34 [17, 49]	224	22	3.7	6653.75	75.5
	RSB-ResNet-50 [17, 49]	224	26	4.1	2732.85	79.8
	RSB-ResNet-101 [17, 49]	224	45	7.9	1856.48	81.3
	RSB-ResNet-152 [17, 49]	224	60	11.6	1308.26	81.8
Attention	DeiT-S [43]	224	22	4.6	3092.02	79.8
	DeiT-B [43]	224	86	17.5	1348.76	81.8
	PVT-Small [48]	224	25	3.8	1622.53	79.8
	PVT-Medium [48]	224	44	6.7	1190.48	81.2
	PVT-Large [48]	224	61	9.8	865.33	81.7
Spatial MLP	MLP-Mixer-B/16 [41]	224	59	12.7	1855.45	76.4
	ResMLP-S24 [42]	224	30	6.0	3228.75	79.4
	ResMLP-B24 [42]	224	116	23.0	298.94	81.0
	Swin-Mixer-T/D6 [26]	256	23	4.0	1625.59	79.7
	Swin-Mixer-B/D24 [26]	224	61	10.4	1131.60	81.3
2D FFT	GFNet-H-Ti [34]	224	15	2.1	1979.56	80.1
	GFNet-H-S [34]	224	32	4.6	1434.19	81.5
	GFNet-B [34]	224	43	7.9	1771.07	80.7
	GFNet-H-B [34]	224	54	8.6	939.20	82.9
Pooling	PoolFormer-S12 [52]	224	12	1.8	4160.18	77.2
	PoolFormer-S24 [52]	224	21	3.4	2140.20	80.3
	PoolFormer-S36 [52]	224	31	5.0	1440.37	81.4
	PoolFormer-M36 [52]	224	56	8.8	1009.45	82.1
	PoolFormer-M48 [52]	224	73	11.6	761.93	82.5
None	RIFormer-S12°	224	12	1.8	4899.80 (↑ 17.8%)	76.9
	RIFormer-S24°	224	21	3.4	2530.48 (↑ 18.2%)	80.3
	RIFormer-S36°	224	31	5.0	1699.94 (↑ 18.0%)	81.3
	RIFormer-M36°	224	56	8.8	1185.33 (↑ 17.4%)	82.6
	RIFormer-M48°	224	73	11.6	897.05 (↑ 17.7%)	82.8
	RIFormer-S12‡	384	12	5.4	1586.51	78.3
	RIFormer-S24‡	384	21	10.0	819.40	81.4
	RIFormer-S36‡	384	31	14.7	552.07	82.2
	RIFormer-M36‡	384	56	25.9	403.15	83.4
	RIFormer-M48‡	384	73	34.1	304.43	83.7

Evaluation

Ablation Study

1) Effectiveness of module imitation

Token Mixer	Feature distillation scheme	Top-1 (%)
Identity	None	74.05
Identity	Feature distill	74.90
Affine	Module imitation	**75.36**

- The accuracy of feature distillation is 0.46% lower than that of module imitation

2) Comparisons of different acceleration strategies

Model	Type	Throughput	Top-1 (%)
PoolFormer-S12	None	4160.18	75.01
PoolFormer-S9	Depth	5025.71	74.78
PoolFormer-XS12	Width	4780.28	75.11
RIFormer-S12	TM	4899.80	**75.36**

- Directly pruning depths or width cannot render a better performance than ours without latency-hungry token mixer

Evaluation

Ablation Study

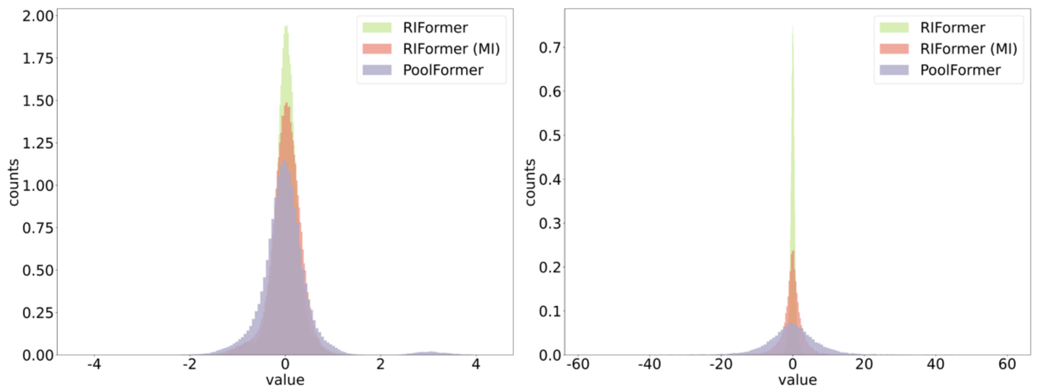
3) Generalization to different teachers

Token Mixer	Teacher	Top-1 (%)
Affine (12 layers)	None	72.75
Affine (12 layers)	RandFormer-S12 [53]	**75.62**
Affine (12 layers)	PoolFormer V2-S12 [53]	**75.87**
Affine (18 layers)	None	75.01
Affine (18 layers)	ConvFormer-S18 [53]	**77.53**
Affine (18 layers)	CAFormer-S18 [53]	**77.26**

Rand matrices, Pooling, Separable Depthwise Convolution, Attention

- Module imitation has a positive effect in different depth setting and teachers

4) Module imitation (MI) shifts the feature distribution of the RIFormer model to be closer to the teacher



(a) Stage 1 (b) Stage 4

- PoolFomer-S12 and RIFormer-S12 show a clear difference in feature distribution.
- The distribution of RIFormer-S12 are basically shifted toward that of the PoolFomer-S12 by module imitation

Outline

- Overview
- Background
- Motivation
- Advantage of Token Mixer-Free Architecture
- A Roadmap
- Evaluation
- Conclusion

Conclusion

- We propose to explore the vision backbone by developing advanced learning paradigm for simple model architecture, to satisfy the demand of realistic application.

- We instantiate the re-parameterizing idea to build a token mixer free vision model, RIFormer, which owns the improved modeling capacity for the inductive bias while enjoying the efficiency during inference.

- Our proposed practical guidelines of distillation strategy has been demonstrated effective in keeping the vision backbone competitive but removing the token mixer.

https://techmonsterwang.github.io/RIFormer/

https://github.com/open-mmlab/MMPreTrain